ONYX Consulting, Inc.

1300 El Paseo, Suite G, PMB 236
Las Cruces, NM 88001-6039

December 5, 2000

Mr. Charles R. Piner
Technical Monitor
US Army Aviation and Missile Command
Attn: AMSAM-RD-WS-DP-SB
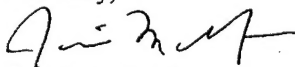Bldg. 7804, Room 222
Redstone Arsenal, AL 35898-5000

RE: Final Status report on SBIR contract DAAH01-00-C-R114

Dear Mr. Piner,

Enclosed please find two copies of the Final Status report for our contract referenced above entitled "Improving Recall in Domain Independent Information". This report covers work performed during the period of April 5 through October 4, 2000. I've also enclosed one completed form DD250 for this report. After your review and acceptance of this report please fax a signed copy of form DD250 back to me at 505-646-6218. I will then proceed with submitting it to DFAS and DCMC it for payment.

Thank you for your assistance with this project. If you have any administrative questions or need additional information please contact me at 505-646-1401. If you have technical questions with respect to the report or the project please contact Mr. Charles Watts, Principal Investigator, at 505-243-1023.

Sincerely,

Jeannine Sandefur
Chief Financial Officer

Encl: DD250
          SF298
          Final Status Report (2copies)

Copy with SF298:   David R. Gunning, ISO DARPA, Arlington, VA
                        Charles Watts, Principal Investigator
                        Commander, US Army Aviation and Missile Command, AMSAM-RD-OB-RC
                        Commander, US Army Aviation and Missile Command, AMSAM-RD-WS
                        Director, Defense Advanced Research Projects Agency, ASBD/SBIR
                        Director, Defense Advanced Research Projects Agency, ASBD/DARPA Library
                        Defense Technical Information Center (2 copies)

IMPROVING RECALL IN DOMAIN INDEPENDENT INFORMATION

Improving Recall For Automatic Information Extraction:
Final Status Report

December 5, 2000

Sponsored by
Defense Advanced Research Projects Agency (DOD)
ISO, David R. Gunning
ARPA Order D611/75

Issued by: **U.S. Army Aviation and Missile Commander**

Contract No.:   DAAH01-00-C-R114

Effective Date of Contract: April 5, 2000

Contract Expiration Date: December 4, 2000

Reporting Period: April 5, 2000 – October 4, 2000

Contractor: **Onyx Consulting, Inc.**

Principal Investigator: Charles Watts

Address: 1300 El Paseo, Suite G, PMB 236
           Las Cruces, NM 88001

Phone: 505-646-1401

REPORT DOCUMENTATION PAGE

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE Dec 5, 2000	3. REPORT TYPE AND DATES COVERED Final Status Report (5 Apr-4 Oct.00)	
4. TITLE AND SUBTITLE Improving Recall for Automatic Information Extraction: Final Status Report		5. FUNDING NUMBERS C:DAAH01-00-C-R114	
6. AUTHORS Charles Watts, P.I., James Cowie and Sergei Nirenburg, Consultants			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Onyx Consulting, Inc. 1300 El Paseo Rd., Suite G, PMB 236 Las Cruces, NM 88001		8. PERFORMING ORGANIZATION REPORT NUMBER NYX0002Z	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) US Army Aviation and Missile Command Mr. Charles R. Piner, Tech Monitor Attn: AMSAM-RD-WS-DP-SB, Bldg 7804, Room 222 Redstone Arsenal, AL 35898-5000		10. SPONSORING/MONITORING AGENCY REPORT NUMBER	

11. SUPPLEMENTARY NOTES

Final Status report for SBIR contract.

12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited	12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This report describes the results of the SBIR Phase I research project on improving recall for automatic information extraxtion, carried out by Onyx Consulting, Inc. Current IE systems suffer from a number of limitations, especially with respect to the types and amount of knowledge that they bring to bear on the process of extraction. The current effort works on improving the coverage and quality of proper name recognition but also significantly on enhancing the syntactic and semantic knowledge used in extraction. In addition, we also work on resolving multiple references to the same entity in the text. Preliminary testing suggests that the knowledge-intensive methods contribute to enhancing the recall in information extraction.

14. SUBJECT TERMS Information extraction, recall, syntax, semantics, ontology, text processing.	15. NUMBER OF PAGES 15
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT

NSN 7540-01-280-5500

Computer Generated

STANDARD FORM 298 (Rev 2-89)
Prescribed by ANSI Std 239-18
298-102

MATERIAL INSPECTION AND RECEIVING REPORT

1. PROC. INSTRUMENT IDEN. (CONTRACT)	(ORDER) NO.	6. INVOICE NO./DATE	7. PAGE	OF	8. ACCEPTANCE POINT
DAAH01-00-C-R114		102/00 Dec 05	1	1	D

2. SHIPMENT NO.	3. DATE SHIPPED	4. B/L	5. DISCOUNT TERMS
NYX0002Z	00 Dec 05	TCN	N/A

9. PRIME CONTRACTOR CODE 088G4	10. ADMINISTERED BY CODE 50302A
Onyx Consulting, Inc. 1300 El Paseo Rd. Ste G, PMB 236 Las Cruces, NM 88001	DCMC Phoenix Two Renaissance Square 40 North Central Avenue Phoenix, AZ 85004

11. SHIPPED FROM (If other than 9) CODE FOB: D	12. PAYMENT WILL BE MADE BY CODE HQ0339
See Block 9.	DFAS - Columbus Center West Entitlement Operations PO Box 182381 Columbus, OH 43218-2381

13. SHIPPED TO CODE W31P4Q	14. MARKED FOR CODE W31P4Q
Mr. Charles R. Piner, Attn: AMSAM-RD-WS-DP-SB US Army Aviation and Missile Command Redstone Arsenal, AL 35898-5000	Mr. Charles R. Piner

15. ITEM NO.	16. STOCK/PART NO. DESCRIPTION (Indicate number of shipping containers - type of container - container number.)	17. QUANTITY SHIP/REC'D*	18. UNIT	19. UNIT PRICE	20. AMOUNT
0001	AB SERVICE: Research and Development Final Status report on "Improving Recall in Domain Independent Information" project.	1	LO	$44,849.00	$44,849.00
	TOTAL:				$44,849.00

21. CONTRACT QUALITY ASSURANCE		22. RECEIVER'S USE
A. ORIGIN ☐ CQA ☐ ACCEPTANCE of listed items has been made by me or under my supervision and they conform to contract, except as noted herein or on supporting documents.	**B. DESTINATION** ☐ CQA ☐ ACCEPTANCE of listed items has been made by me or under my supervision and they conform to contract, except as noted herein or on supporting documents.	Quantities shown in column 17 were received in apparent good condition except as noted.
DATE SIGNATURE OF AUTH GOVT REP	DATE SIGNATURE OF AUTH GOVT REP	DATE RECEIVED SIGNATURE OF AUTH GOVT REP TYPED NAME AND OFFICE
TYPED NAME AND OFFICE	TYPED NAME AND TITLE	* If quantity received by the Government is the same as quantity shipped, indicate by (˅/mark; if different, enter actual quantity received below quantity shipped and encircle.

23. CONTRACTOR USE ONLY

Onyx Consulting, Inc. Point of Contact: Jeannine Sandefur   505-646-1401 (FAX 505-646-6218)

DD FORM 250, NOV 92 (EG)   Previous edition may be used.   Designed using Perform Pro, WHS/DIOR, Nov 95

Improving Recall for Automatic Information Extraction
December 5, 2000

Final Report

Abstract

This report describes the results of the SBIR Phase I research project on improving recall for automatic information extraction, carried out by Onyx Consulting, Inc.

Current IE systems suffer from a number of limitations, especially with respect to the types and amount of knowledge that they bring to bear on the process of extraction. The current effort works on improving the coverage and quality of proper name recognition but also significantly on enhancing the syntactic and semantic knowledge used in extraction. In addition, we also work on resolving multiple references to the same entity in the text. Preliminary testing suggests that the knowledge-intensive methods contribute to enhancing the recall in information extraction.

Subject Terms

Information extraction, recall, syntax, semantics, ontology, text processing.

Background

Current information extraction systems work basically as follows. A set of *key words and patterns* is developed for a domain (e.g., nuclear physics or reports about terrorist activities). Each word signals and corresponds to an information element (such as name, event-type or place) which is sought in the extraction process. The patterns are used to detect appropriate combinations of words and names. In practice, the information elements recognized by such patterns are organized in *templates*. The key words are used as *textual clues* in that they are matched against texts from a collection. Those documents in which matches occur are further processed to find the patterns which allow the filling of the slots corresponding to the various elements of the pattern on which the match occurred.

Our recall enhancing algorithms are based on using additional knowledge in the process of information extraction; additional, that is, compared with most current systems. In particular, we are adding a combination of syntactic and ontological-semantic knowledge to support template definition and the extraction process itself. Our algorithms are built to allow the exploration, in combination, of a number of approaches to breaking the recall barrier. Resources and linguistic expertise were supplied by consultants at the Computing Research Laboratory, New Mexico State University.

General Progress

Onyx Consulting has fulfilled the objectives of this Phase I project—to demonstrate that levels of recall in information extraction can be raised through the application of natural language processing techniques, centrally including analysis of the text meaning. In what follows, we

briefly describe the design, algorithms and data resources of the proof-of-concept application system that we built during Phase I.

The proof of concept system that we built, while basically an information extraction system, is presented to the user in the form of a question answering environment. Since the Chechnya war has somewhat lost its newsworthiness, we decided to use a database of news about the Sydney Olympic Games. Since the domain is not, at this stage of development, the most important consideration, in our choice we were guided, in part, by the ready availability of textual material about this event.

Figures 1 - 9 illustrate the way the proof-of-concept system works.
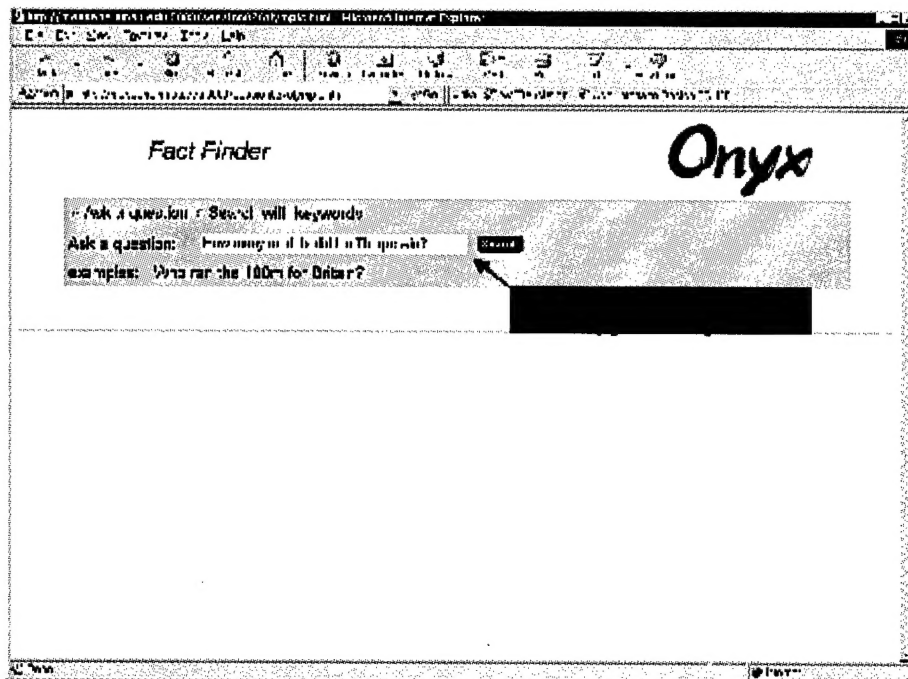
Figure 1. The user types in a question, for example, "How many medals did Ian Thorpe win?"
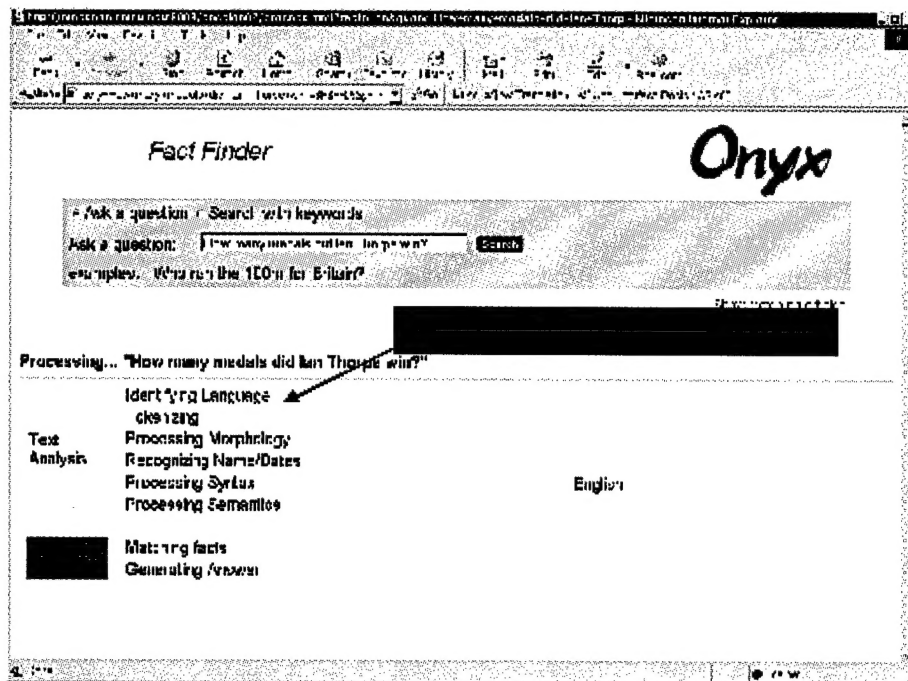
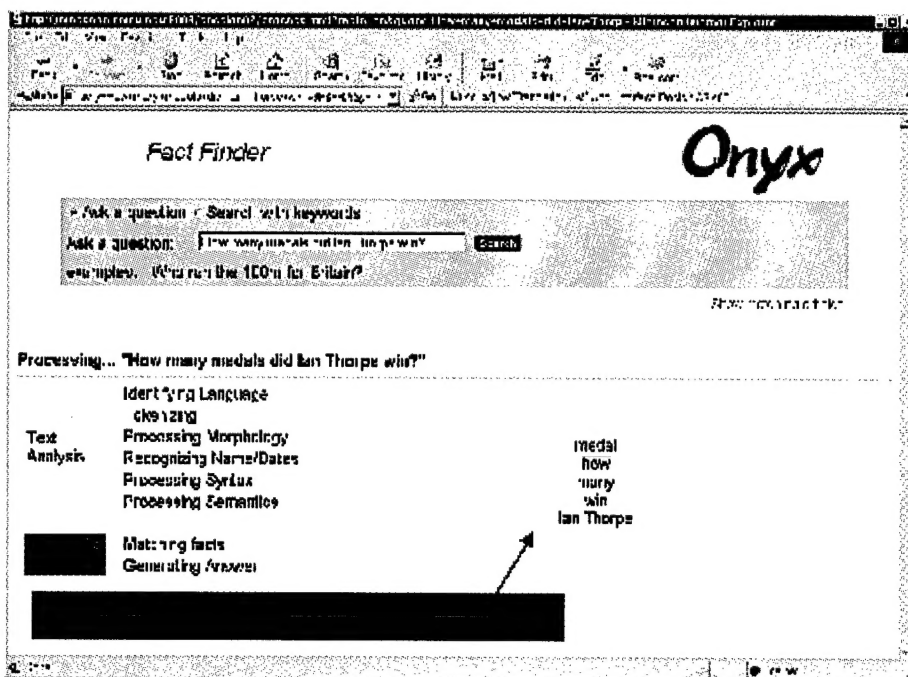**Figure 2.** The system starts a battery of processing stages, as listed in the display

Figure 3. The results of each stage are displayed. This is a testing/debugging feature used in the proof of concept that will be only an option in a Phase II prototype.
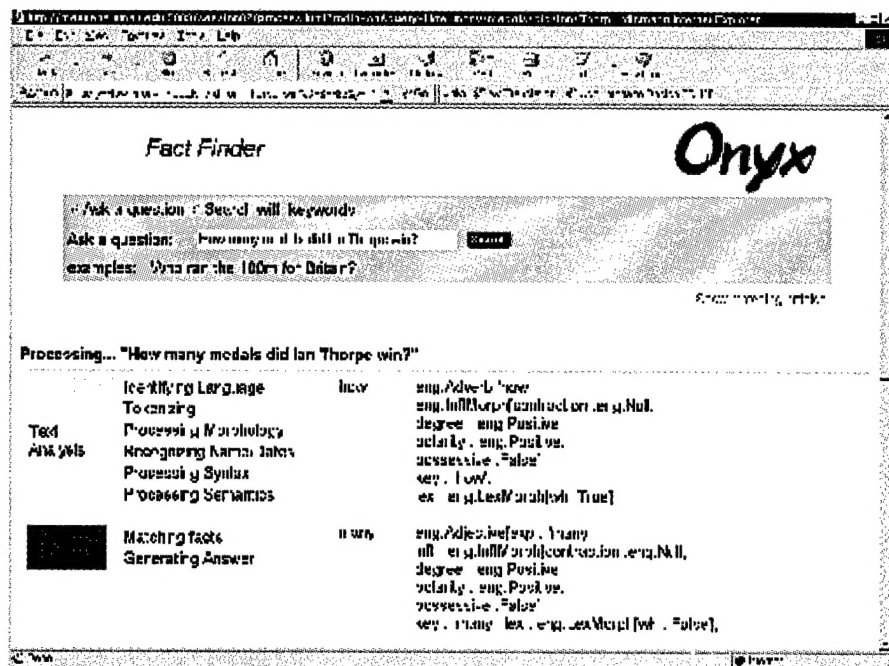
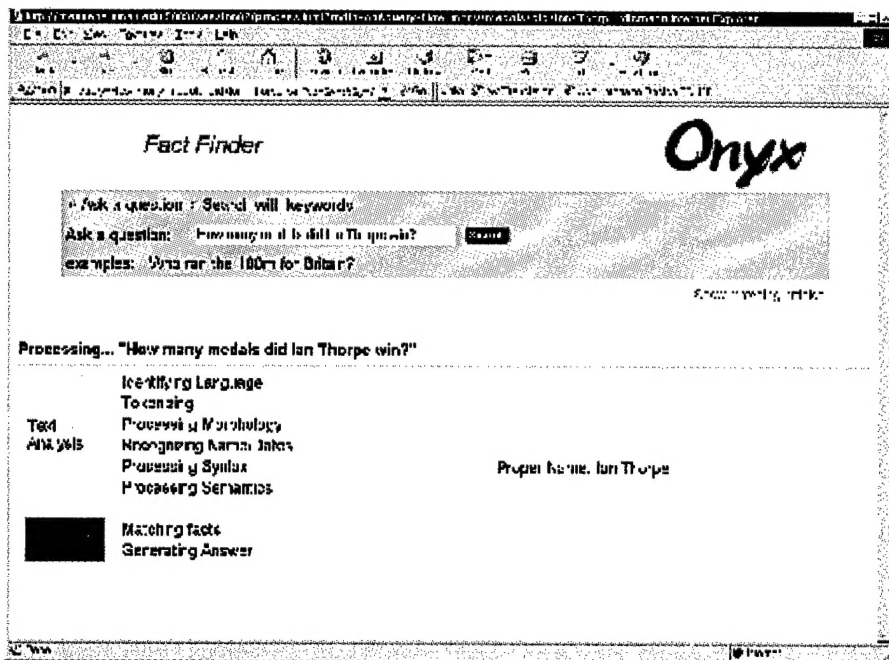**Figure 4**. The results of morphological analysis of the input string are displayed.

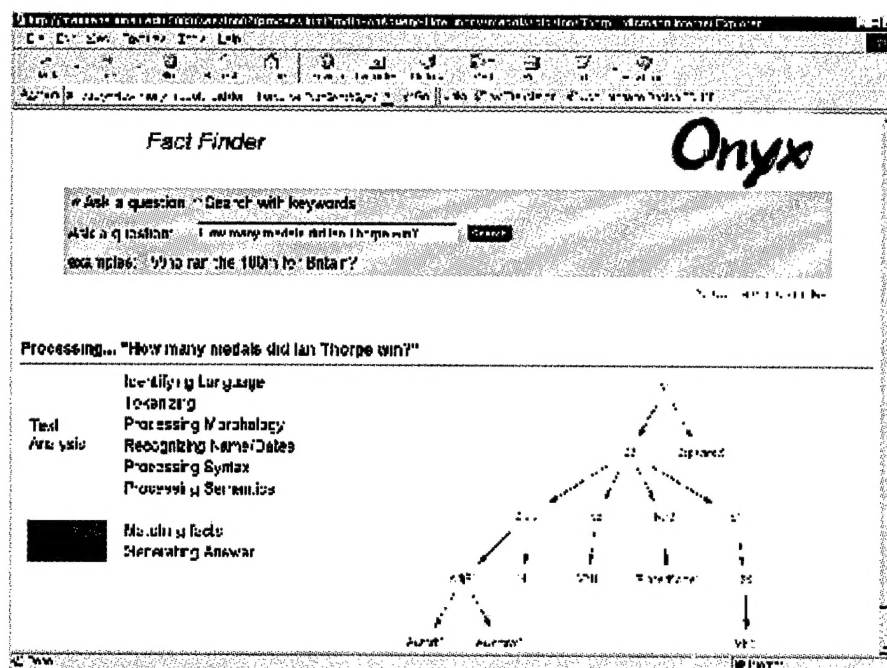**Figure 5**. The results of proper name recognition are displayed.

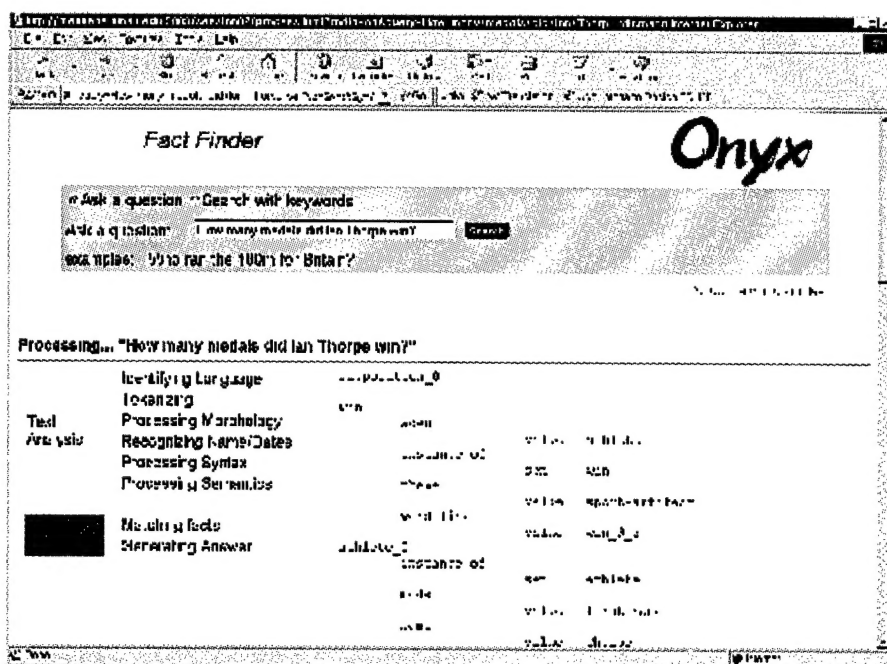**Figure 6**. The results of syntactic analysis are displayed graphically.

Figure 7. The results of semantic analysis are encoded in a format compatible with the format in which facts are stored in the Fact DB and, therefore, with the format in which ontological knowledge is encoded.
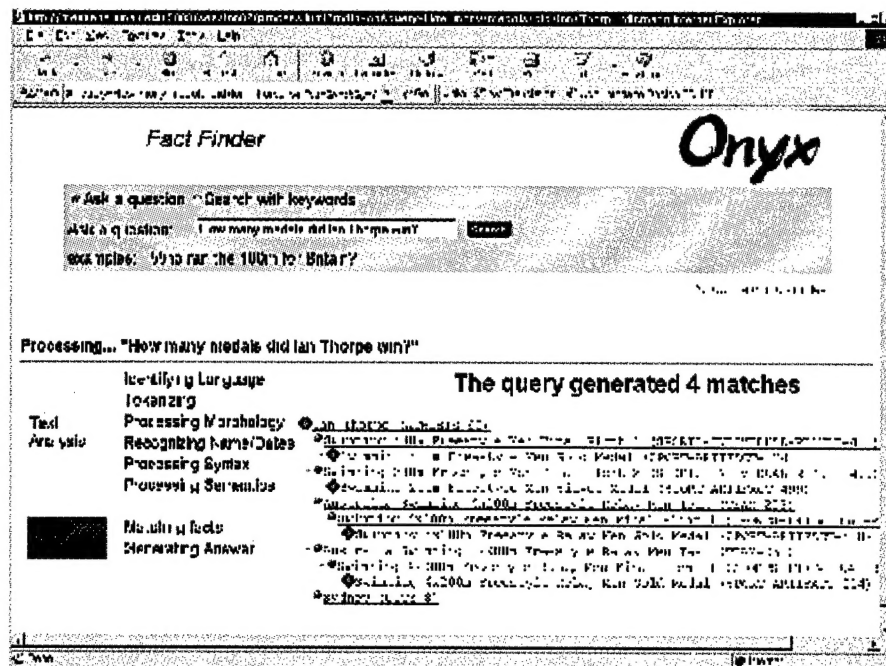
Figure 8. A view of a dynamically produced hierarchy of facts from the Fact DB that will form the answer to the query.
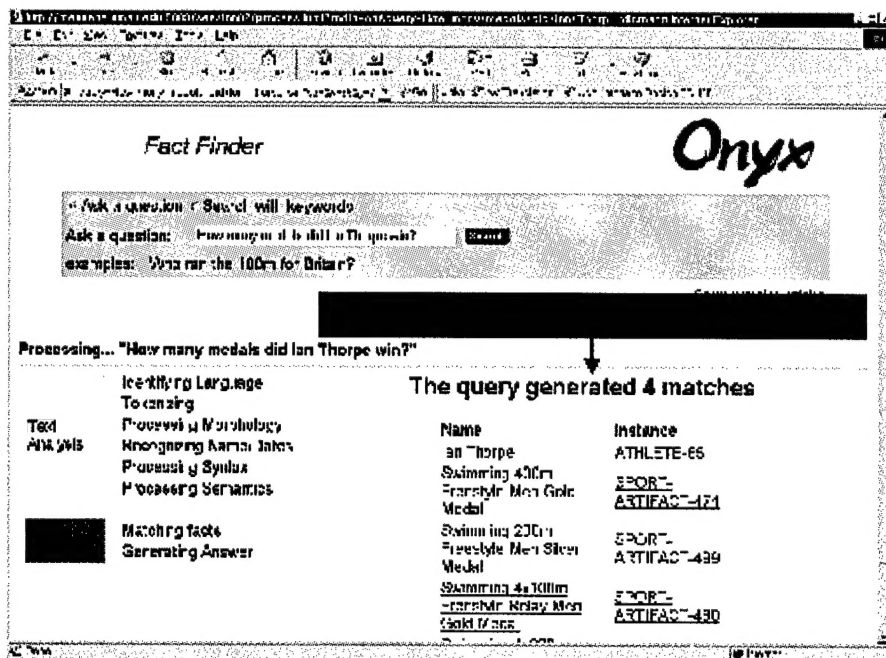
Figure 9. This is how the final results of the query are displayed.

The processing that is going on behind the scenes at the stage after the battery of input analysis programs (stages) have completed their work is illustrated in Figures 10 - 13, using the same sample query.
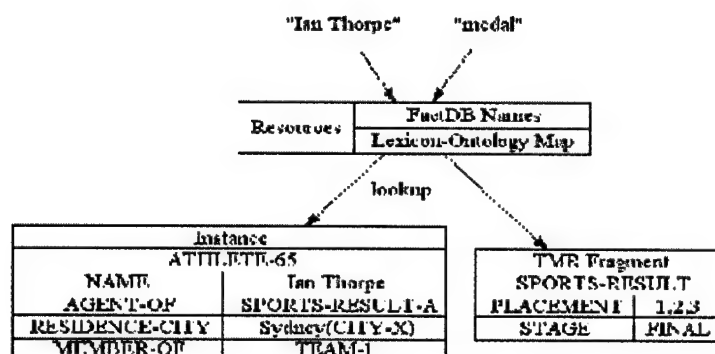
Lookup through Fact DB and Lexicon

Figure 10. The fact database links information structures representing the various elements of the question. In an extraction system this can be used to confirm relationships which are ambiguous in a text.

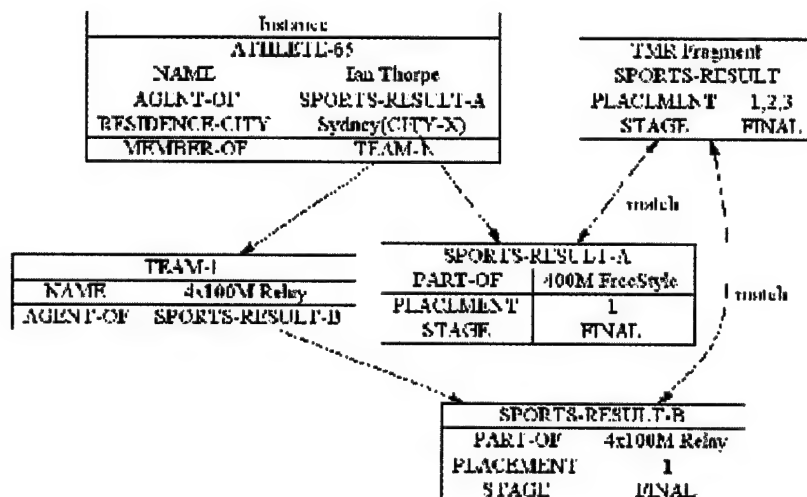## Finding shallow/deep matches through expansion...

Figure 11. Finding links between question elements. The structures here are the product of information extraction, with a human editor in the loop.

Following false paths:

- Don't follow inverse links
- Validate connections
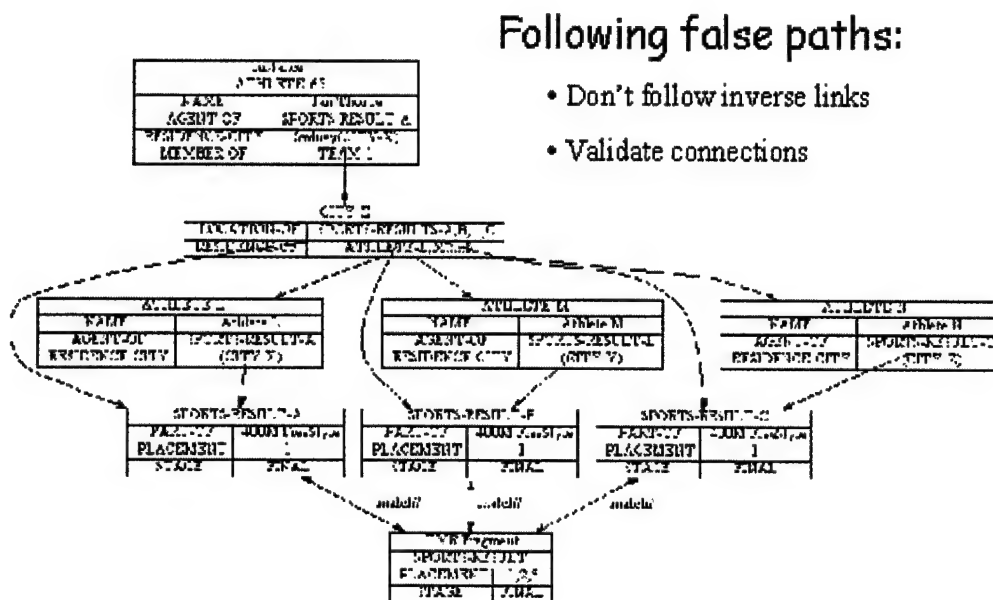
Figure 12. Heuristics are applied to avoid the generation of spurious links in the fact DB.
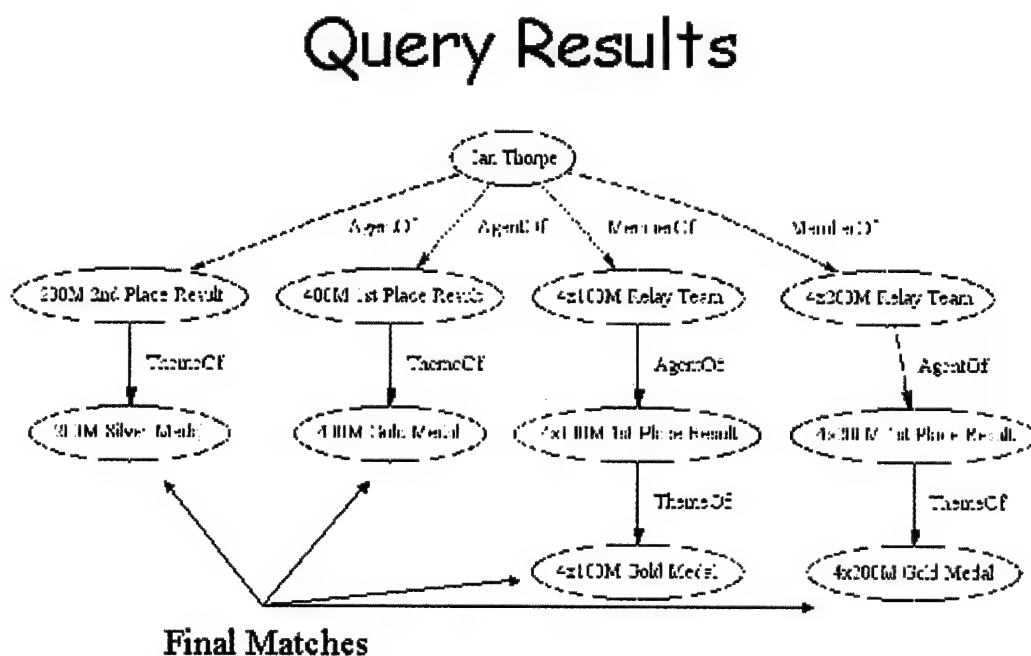
Query Results

Final Matches

Figure 13. The final facts which supply the query answer are selected.

Coverage of Original Tasks

In what follows, we briefly describe the results that we obtained with respect to the five original tasks specified in the Statement of Work of Phase I of this project.

Task 1: Name recognition and classification capabilities

The two-pass name recognition algorithm allowing for the concurrent consideration of multiple name patterns has been implemented. The first pass is a basic lookup and text tokenization stage and the second pass is a bottom-up pattern matching algorithm, which allows multiple instances of names in a text to be recognized by the same pattern application step. The current onomasticon database contains about 200,000 entries including countries, cities, companies and human first and last names. This resource has been incorporated into the new name recognition algorithm.

Task 2: Treating part-of-speech ambiguity

A new syntactic grammar of English has been developed at CRL and tested by Onyx. The syntactic analyzer relies on a locally developed lexicon-based part of speech tagger. The syntactic phase is closely coupled with the name recognition and classification phase (Task 1). The syntactic analyzer called MEAT, developed at CRL, has been adapted for this task.

Task 3: Integration of ontology and ontological-semantic lexicon

A new version of the ontology including browser and editor tools has just been made available to Onyx. The onto-search word disambiguation software has been rewritten (from Lisp) to provide a phrase and sentence level set of disambiguation. It became clear that developing a completely new domain area for the testing of this system was not feasible given the available resources. Therefore, instead of the domain of the Chechnya conflict, Onyx used one of the domains developed at CRL, namely, the domain of Olympic Games. A working version of the extractor, including disambiguation without reference resolution, has been developed.

Task 4: Treatment of reference

The first version of the reference resolution module has been developed and tested. However, the results and the coverage of reference-related phenomena have not yet attained an acceptable level. Thus, the reference module to date concentrates exclusively on anaphoric pronominal co-reference, to the exclusion of treatment of definite descriptions, ellipsis and other means of co-reference expression. The work on expansion of coverage of the reference treatment module will be included in the work on Phase II of this project.
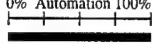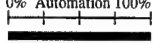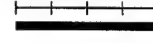
Task 5: Build and test integrated system

While working on the integrated system, as illustrated above, we concentrated on the software engineering tasks of integrating the various system modules. It became clear in the process of work that the only evaluation regimen that was feasible under the constraints of the available resources was an informal qualitative one.

As a result of our work in Phase I, we have arrived at the following evaluation of current capabilities and immediate prospects for enhancement of these capabilities, in terms of

enhancing the levels of automation of the various processes involved in our information extraction / question answering environment. Figures 14-16 illustrate our findings and estimates.
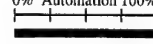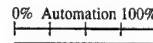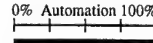
NL Input → NL Output
Processes

Process	Resources	Automation
Language Recognition	Algorithms, Training Data	0% Automation 100%
Tokenization, Morphology	Lexica	0% Automation 100%
Name & Date Recognition	Onomastica, MITRE	0% Automation 100%

Key(s): Current Automation Level

Figure 14.
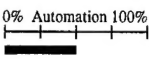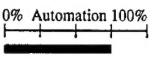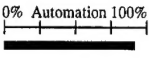
NL Input → NL Output
Processes

Process	Resources	Automation
Syntactic Analysis	Lexica, Grammars, Onomastica	0% Automation 100%
Semantic Analysis	Ontology, Lexica	0% Automation 100%
Answer Generation	Ontology, Lexica, FDB	0% Automation 100%

Key(s): Current Automation Level, Attainable Automation Level,
      FDB = Fact Database

Figure 15.

Resource Acquisition
Processes

Process	Resources	Automation
Ontology Acquisition	Raw & Processed Data	0% Automation 100%
Lexicon Acquisition	Ontology, Raw & Processed Data	0% Automation 100%
Fact Acquisition	Ontology, Lexica, FDB, Web, Data	0% Automation 100%

Key(s): Current Automation Level, Attainable Automation Level,
FDB = Fact Database

Figure 16.

Further Improvement of Recall Levels

Phase I of this project has resulted in a proof-of-concept system that improves both precision and recall characteristics of current IE systems. Thus, the inclusion of syntactic and semantic analysis features enhances the levels of precision by ruling out those part-of-speech and meaning homographs of the terms in the original query that were not intended by the user. As to the level of recall, we established that it can be enhanced in a system that can take into account any or all of the following considerations:

- when the system is aware of the different ways of referring to a certain entity, specifically, through

    o direct mention (e.g., US President),

    o an expression from the synonymy set of direct mentions (President of the United States, the President, Mr. President [in direct speech], etc.),

    o a pronominal reference (e.g., he), a reference by name (e.g., Roosevelt),

    o a reference by ellipsis (as in the second clause of the following sentence: *Roosevelt opposed isolationism and went on to win the 1940 election in part on this platform*),

    o a reference by definite description (as in *the occupant of the Oval Office*)

    o a reference using non-literal language (as in *The White House*)

- when the system is capable of picking out from running text references to events and objects that have a direct causal or other proximal relationship to the concept in the query; for example, if a question is about bankruptcies, relevant information may also appear in company earnings reports (the system must know from its ontology that low -- or lower than expected -- earnings may eventually lead to bankruptcy); as another example, a political demonstration may be classified in some news source as a protest event (which in an ontology would be a parent of the demonstration concept); such causal, temporal and hierarchical links among various events and objects are the main content of a realistic ontology, and thus will help in raising recall by finding texts that, while not directly mentioning the events or objects expected in the original template identified as the most relevant one for a particular query, activate closely related templates.

Commercialization Need and Global Potential

While the basic technology under development under this project has been essentially concentrating on English, Onyx is in a very good position to rapidly apply it to many other languages as well as to text collections in multiple languages.

Recent estimates of total pages on the Internet exceed one billion. Knowledgeable sources indicate that as many as a million new pages are added each day. Although much of the initial

activity of e-commerce and rapid development of the World Wide Web occurred in the United States and other English-speaking countries, current activity has spread at lightning speed to the entire world.

Ninety-two (92) percent of the world's population are non-native English speakers. Idiom, Inc., a leader in e-commerce globalization, quoting a number of reliable sources, estimates that by 2002, non-English speaking users will make up over 50% of the total online population, and 70% by the year 2004. (International Data Corporation). Non-native English speakers make up the fastest growing group of Internet users. (New York Times). Business Web users are three times more likely to buy when addressed in their own language. (Forrester Research). The global e-commerce market is forecasted to grow from $13 billion in 1997 to $1.2 trillion by 2001. (Coopers & Lybrand). Non-US e-commerce will shift from 14% of the total worldwide revenues to 37% by 2002. (International Data Corporation). By the year 2002, 490 million people around the world will have Internet access, that is 79.4 per 1,000 people worldwide, and 118 people per 1,000 by year-end 2005. The top 15 countries will account for nearly 82 percent of the worldwide Internet (The Computer Industry Almanac).

Automated information retrieval and extraction, including that from documents in multiple languages, will be the *sine qua non* of a competitive edge in the increasingly globalized world economy. As in military and intelligence operations up to now, having the best information the soonest will be a critical factor in either gaining increased market share, profitability and 'winning' or becoming an 'also ran,' succumbing to the predations and superior technological capabilities of competitors.

The current project is an important step toward the next generation of information extraction engines. Initial discussions with venture capital brokers suggest that interest in further development of this kind of software is significant. As an 'add on' to existing search engines or as a standalone, customizable to particular products, markets and languages, the enhanced information extraction capability Onyx is developing will make it an extremely desirable, if not essential, tool for any vendor interested in tapping the growing international market of business to business commerce.

In addition, multilingual data mining and text analysis to support market research and new product development activities, as well as direct sales to individual consumers, other businesses and governmental entities worldwide, will be significantly enhanced through the use of a multilingual interactive document summarizer and information retrieval engine. The enhanced-recall, possibly multilingual software product requires additional research and development, supported with SBIR funding in Phase II, and with partners from private industry and venture capital sources, to reach the next level of information extraction capability required for profitable product commercialization. We believe that Phase I has corroborated our belief in the feasibility of the approach. What remains is the additional development necessary to bring the software, and it customization capabilities, to the global market.

The Business Model

We believe that the technology under development in this project has direct and immediate business applicability. At this state in the development of natural language processing

technology it is not realistic to expect fully automatic processing of large quantities of written and spoken language in unconstrained domains. Organizations that bank on such capabilities have traditionally and routinely failed, often -- as in the case of the Belgian company Lernaut & Hauspie, spectacularly so, after a period of unconstrained, though misguided, growth. We propose a more realistic business model, in which Onyx will offer the various organizations in need of improving the coverage, throughput and/or quality of their text processing tasks a service that will offer the following options:

- Onyx learns the domain (or domains) of interest to a client and develops the ontology / lexicon / onomasticon / fact DB resources necessary for enhancing the quality of text processing in these domains; supplies the IE system with this newly acquired knowledge and installs this capability on the client's system; and offers technical support and maintenance;

- Onyx trains the client's representatives in the art of acquiring and maintaining the necessary knowledge resources and delivers the IE system together with a knowledge elicitation environment (based on the Boas language acquisition environment developed at CRL -- see, e.g., Nirenburg and Raskin 1998 -- and available to Onyx) to facilitate the knowledge operations by the client's personnel. As in the previous option, Onyx also offers continued technical support and maintenance.

Irrespective of which of the above avenues is chosen, the type of system that Onyx will develop for the clients will depend also on security considerations and such parameters as development time and funds available and the desired level of throughput. The areas in which our technology offers the most impact include: economics reporting, intelligence message traffic, processing legal texts, automatic abstracting of research production and corporate archive maintenance and use.

Despite the growth of internet companies basing business models on provision of web services, there is an emerging problem of profitability of these enterprises - people expect internet data to be free of charge. However, for many of the areas cited above accurate, and complete (high recall data) is at a premium. Onyx plans to develop a service based model, using the technology of information extraction, information retrieval, question answering, and machine translation used by trained support staff to provide a high recall, high precision data source for our clients' immediate needs in any context.